Interpreting Fine-grained Dermatological Classification with Deep Learning

S Mishra [1], H Imaizumi [2], T Yamasaki [1]

1The University of Tokyo
2ExMedio Inc

ISIC Skin Image Analysis Workshop
Scope

• Analyze model accuracy gap on benchmark datasets (CIFAR-10) vs. dermatological image corpus (DermAI*)
  • SOTA on CIFAR ~98%, whereas dermoscopic ~90%

• Investigate leading label pairs by case studies
  • 3 leading pairs investigated by GradCAM/GBP

• Suggestions on better datasets of user-submitted images by our experience
  • Data Augmentation, FoV, Gamma & Illumination correction
Dataset

User submitted Dermoscopic images across 10 most prevalent labels. 7264 images, split in 5:1 (train/test)

- Acne
- Alopecia
- Blister
- Crust
- Erythema
- Leukoderma
- P. Macula
- Tumor
- Ulcer
- Wheal
Dataset

• Addressing the most common dermatological complaints.

• Ultimate goal: To perform reliable rapid screening to reduce outpatient burden.
Model Learning

• Test several architectures of increasing size/complexity

• 5:1 split, Early stopping, BCE with logits loss
  • Learning rate range test
  • SGD + Restarts (SGD-R)
    • SGD-R + Length Multiplication+ Differential Learning

• Modus operandi tested on CIFAR-10 prior*
Steadily increase the LR and observe the Cross entropy loss Test several mini-batches to see a point of inflexion

Reference:
Cyclical Learning rates for training NN, L. Smith [2017]
Deep Learning, S. Verma et al. 2018
SGD-R

1. Avoid monotonicity by Cosine scheduling function

\[ v(t) = \frac{1}{2} \left( 1 + v \cos \left( \frac{t \pi}{T} \right) \right) + \epsilon \]

2. Cycle Length Multiply by integral powers of 2 over whole architecture

Reference:
SGD with Warm restarts, Loschilov [2017]
# Application

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Acc. (Top-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-34</td>
<td>88.9%</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>89.7%</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>88.2%</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>89.8%</td>
</tr>
</tbody>
</table>

ResNet 152 Confusion Matrix
Analysis

• Following best practices still leaves gap.
• Focus on the label pairs which account for most errors.
• Use GradCAM and Gradient Backprop to analyze what CNNs capture in learning process.

<table>
<thead>
<tr>
<th>Label 1</th>
<th>Label 2</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ulcer</td>
<td>Tumor</td>
<td>29</td>
</tr>
<tr>
<td>Macula</td>
<td>Erythema</td>
<td>25</td>
</tr>
<tr>
<td>Blister</td>
<td>Erythema</td>
<td>17</td>
</tr>
<tr>
<td>Erythema</td>
<td>Wheal</td>
<td>15</td>
</tr>
<tr>
<td>Crust</td>
<td>Ulcer</td>
<td>14</td>
</tr>
<tr>
<td>Blister</td>
<td>Crust</td>
<td>14</td>
</tr>
<tr>
<td>Macula</td>
<td>Tumor</td>
<td>13</td>
</tr>
<tr>
<td>Macula</td>
<td>Leukoderma</td>
<td>10</td>
</tr>
<tr>
<td>Blister</td>
<td>Ulcer</td>
<td>7</td>
</tr>
<tr>
<td>Tumor</td>
<td>Erythema</td>
<td>7</td>
</tr>
<tr>
<td>Crust</td>
<td>Tumor</td>
<td>5</td>
</tr>
</tbody>
</table>

Label pairs with at least 5 errors

Reference:
GradCAM: Visual explanation from DNN, Selvaraju [2016]
Guided BP, Springenberg [2014]
### Ulcers & Tumors

<table>
<thead>
<tr>
<th>Ulcer 0.391</th>
<th>Tumor 0.152</th>
</tr>
</thead>
<tbody>
<tr>
<td>High degree of geometrical (spherical) similarity is the common factor in many samples</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tumor 0.78</th>
<th>Ulcer 0.212</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevations and inflammations seen in Tumors, misclassifies many ulcer samples.</td>
<td></td>
</tr>
</tbody>
</table>
Macula & Erythema

<table>
<thead>
<tr>
<th>Erythema 0.53</th>
<th>Macula 0.41</th>
</tr>
</thead>
</table>

Presence of pigmentation patches around the lesion can mispredict.

*FoV and ROI selection could lead to better results.*

<table>
<thead>
<tr>
<th>Macula 0.69</th>
<th>Erythema 0.28</th>
</tr>
</thead>
</table>

Oval/cycloidal patches makes GBP confused with the overall shape of Macula.

FOV & Depth important factors to consider
Ulcer & Crust

<table>
<thead>
<tr>
<th>Crust 0.86</th>
<th>Ulcer 0.124</th>
</tr>
</thead>
</table>

Presence of large centroid is possible source.

*Difficult to predict as both related chronologically*

<table>
<thead>
<tr>
<th>Ulcer 0.91</th>
<th>Crust 0.06</th>
</tr>
</thead>
</table>

Oval/cycloidal patches on GBP

*Selection of right RoI, illumination could improve many cases.*
Mitigation

Highlight some of the “hard-learned lessons” building this project from scratch.

Mitigation factors to look out:

- Balancing training sets (dynamic vs static)
- Field of View / ROI selection
- Illumination and Gamma correction
Custom datasets can be small, unevenly divided. Best to use dynamic in-memory augmentation during batch selection. Larger batches preferably.
Field of View/Object Depth

FOV selection dramatically improves performance. In user-submitted images, pre-processing needed. Bonus: if illumination stable
Gamma & Illumination

Often illumination & shadow effects

Gamma adjustment $\approx 1.2 – 1.5$

Creating illumination map & reversing imbalanced lighting by normalizing.

Prediction: Ulcer (98%)
Actual: Tumor (1%)

Prediction: Tumor 78%
Conclusion

• Gap may never be entirely removed,

• [Status Quo] Racial diversity one of the hardest problems to crack. Better to focus on single one for better performance. (But harder in developed countries).

• Not all artifacts can be fixed in user-submitted images.

• Augmentation & Photo-grammatic corrections can improve the quality of model learning/inference dramatically.
  • Balancing training data, FOV reduction, Gamma & illumination correction
Thank you!
Scope

Rapid improvements in image classification tasks

• Larger better & detailed datasets
• Faster hardware resources
• Better architectures

However (the ugly truth)!
• More iterations to SOTA
• Longer train time
• Higher costs
• Small dataset reliability low
Scope

Deployment costs can adversely impact individuals or smaller groups.

SOLUTION?

• Organic combination of proven techniques, field tested on benchmark datasets.
• Optimization by learning rate \((\nu)\) adaptations.
• Transfer modus-operandi to smaller, untested data.
• Ensure repeatability.
CIFAR Baseline

• Multi-class classification on CIFAR-10
• Test candidate architectures of increasing size/complexity
  - DenseNet161
• Baseline Performance
  5:1 split, Early stopping, lower LR restarts
  BCE with logits loss
  Train to 90%+ validation accuracy mark
Differential learning

Gear-box need not spin all gears equally!

Reduce computational overhead by assigning different learning rates.

Courtesy: J Howard, T. Parr [2018]
## CIFAR Baseline

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Accuracy (Top-1)</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet 34</td>
<td>90.36%</td>
<td>17,757</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>90.54%</td>
<td>34,039</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>90.71%</td>
<td>60,639</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>90.68%</td>
<td>91,888</td>
</tr>
<tr>
<td>DenseNet-161</td>
<td>93.02%</td>
<td>54,628</td>
</tr>
</tbody>
</table>
## CIFAR Speedup Results

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Accuracy (Top-1)</th>
<th>Time (s)</th>
<th>$\eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet 34</td>
<td>96.84%</td>
<td>9,565</td>
<td>1.84</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>96.82%</td>
<td>11,817</td>
<td>2.88</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>97.61%</td>
<td>6,673</td>
<td>9.09</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>97.78%</td>
<td>9,012</td>
<td>10.2</td>
</tr>
<tr>
<td>DenseNet-161</td>
<td>97.15%</td>
<td>7,195</td>
<td>7.59</td>
</tr>
</tbody>
</table>
Higher dividends when architecture size grows larger. Possible by offsetting the computation overhead by DLR
CIFAR Results

DenseNet 161

ResNet 152

* Appendix